Self-deposit, discovery, and delivery of scientific GIS datasets using GeoHydra

Darren Hardy
Stanford University
Open Repositories 2014, Helsinki, Finland

The Story

- Marine ecologist Malin Pinsky
- Published research on Pacific salmon conservation
  - Article: Pinsky et al. 2009, *Conservation Biology* 23(3)
  - Conducted while at Stanford's Hopkins Marine Station
  - Used in testimony before the United States Senate, Apr 2010
- Self-published GIS data on his personal website
- Graduated, left Stanford for Rutgers, Assistant Professor
- Website taken down(!)… 404 Not Found
- Self-deposited his research data into Stanford repository
  - Now, discovery, delivery, and preservation services

Self-publish vs. Self-deposit

- Self-publish
  - Personal or departmental website, or email ☺
  - Typically, delivery services only

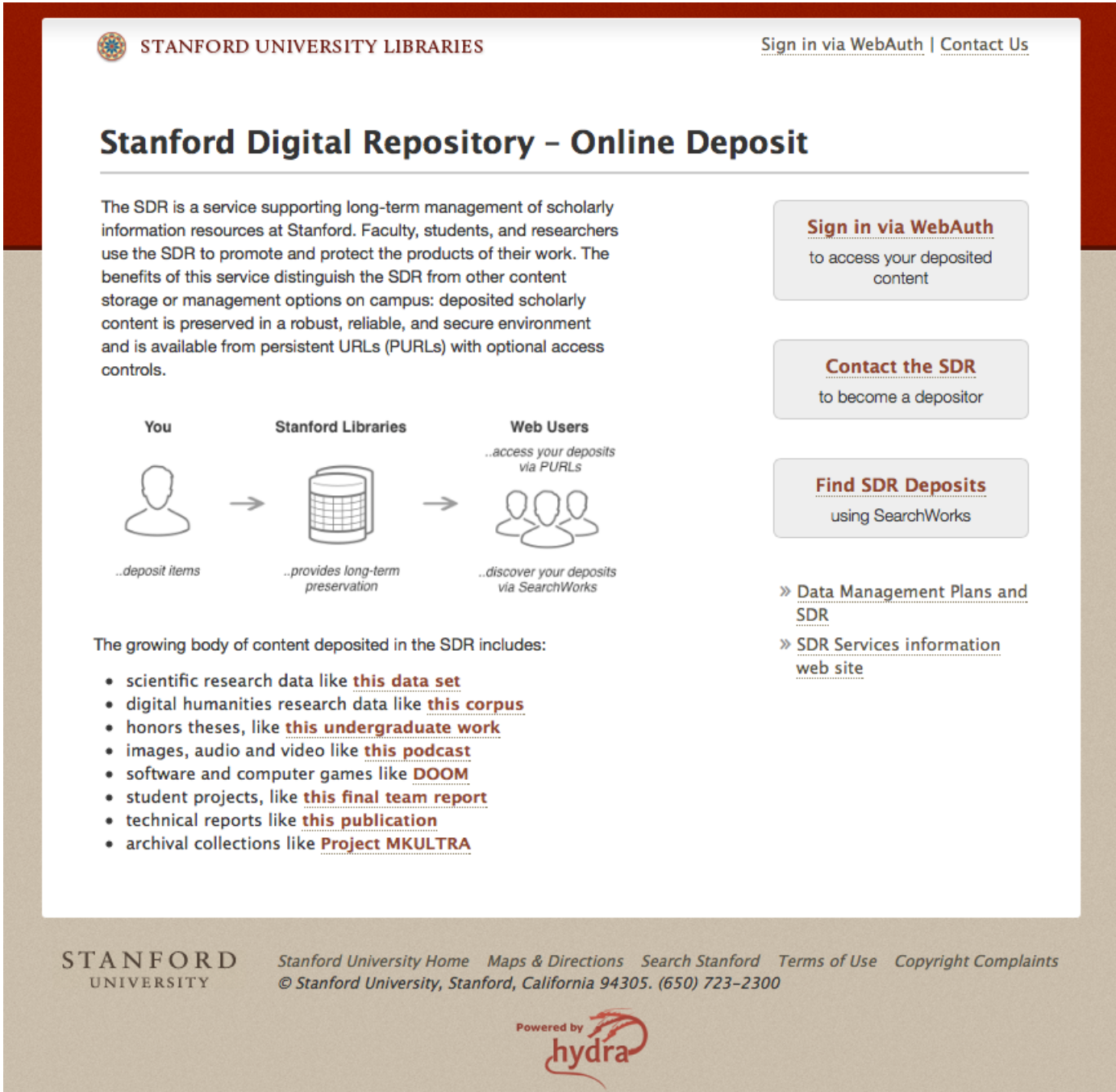
- Self-deposit
  - Upload file(s), provide bibliographic metadata
  - Stanford Digital Repository, an institutional repository for scholarly information resources
  - Repository can provide variety of services
    - Discovery, delivery, preservation, curation

SDR as of Jan 2014

453 collections
189 million files
137 TB

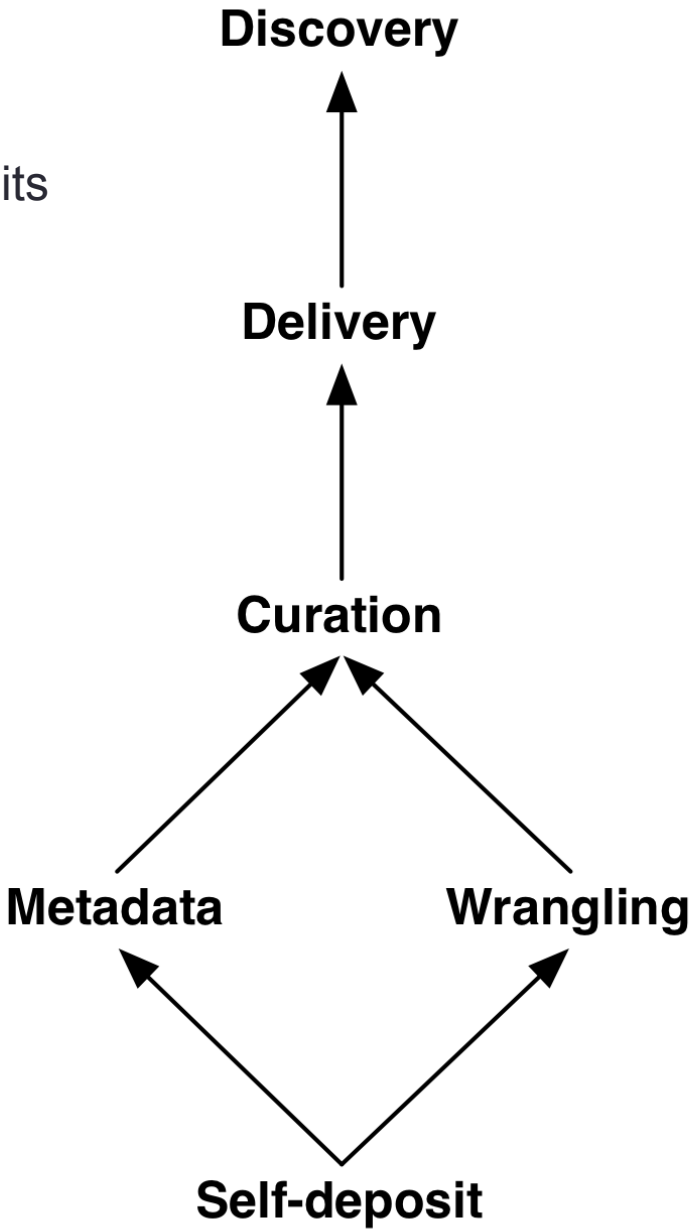
Articles, data sets, theses, photos, media, maps, manuscripts, etc.

See OR14 presentation on SDR by Hannah Frost on Thursday 1115 (7A)



The Approach

Provide services for GIS self-deposits



The GIS Data from Pinksy

- Biological measure of conservation value
- Six species of Pacific salmon (*Oncorhynchus spp.*)
- ~800 catchments of the northern Pacific
  - Canada, China, Japan, Russia, and the United States

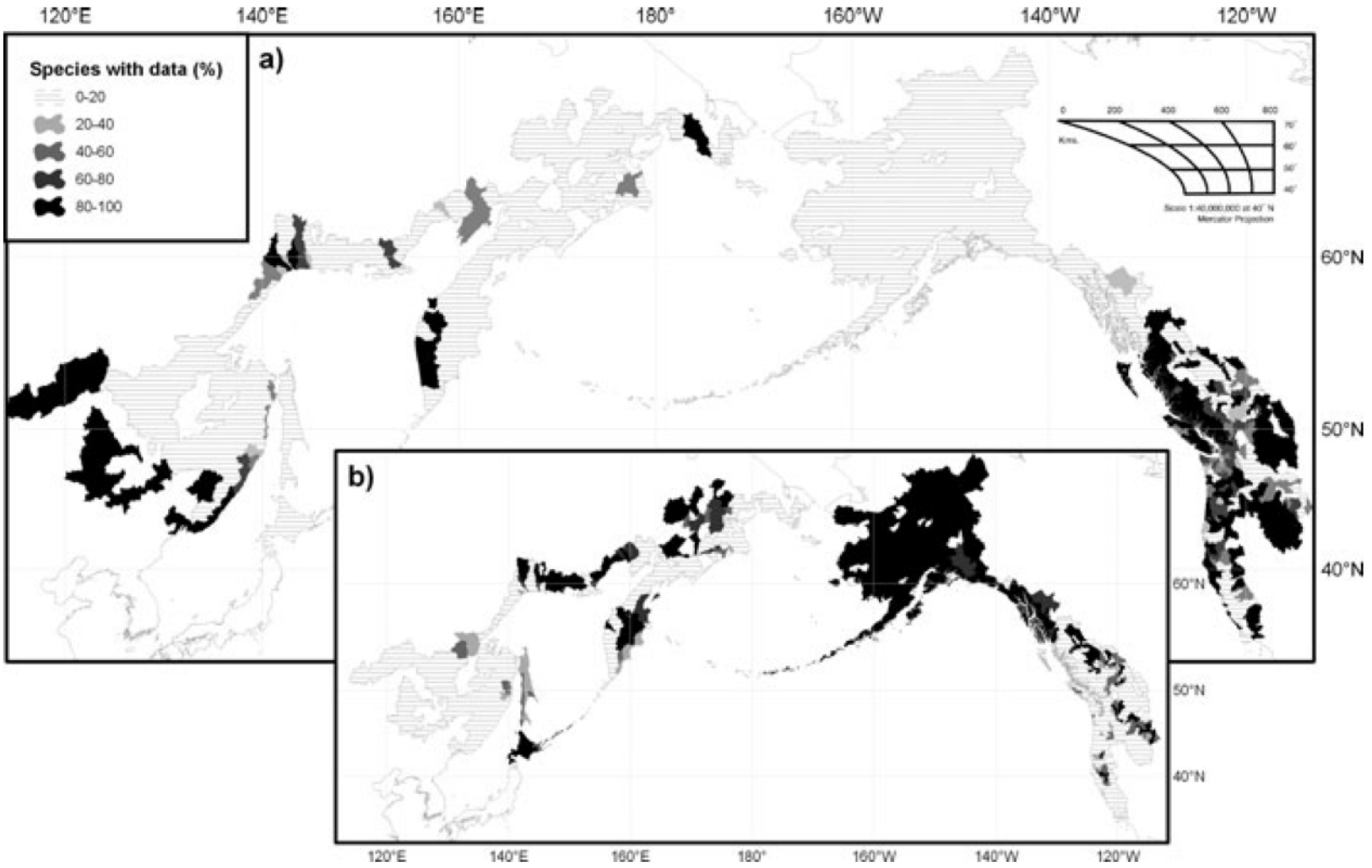
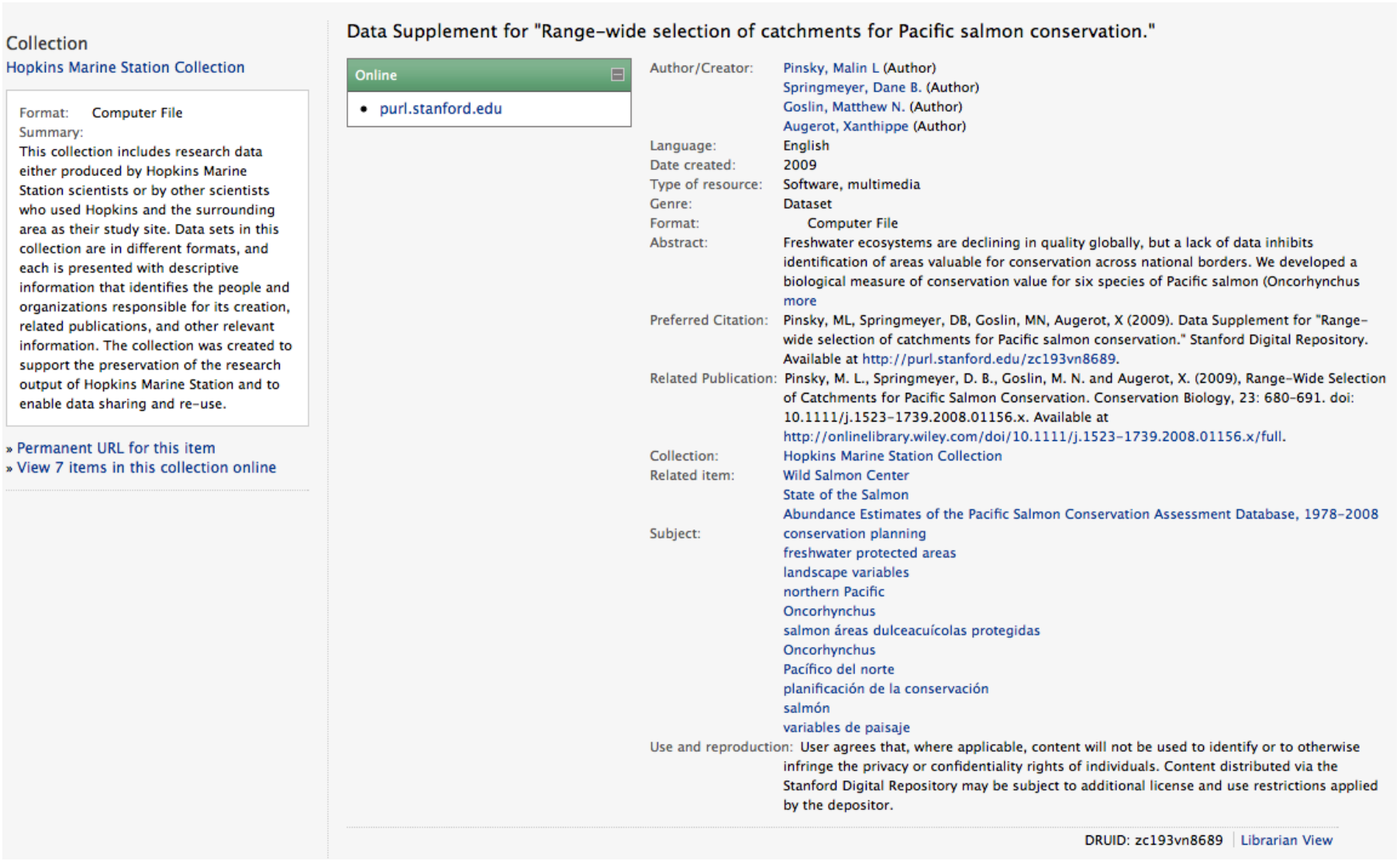


Figure 1. Quality of salmon abundance data used to calculate conservation value (CV), displayed as the percentage of species in each catchment for which data of a certain spatial scale were available: (a) data available at the scale of individual catchments or smaller and (b) data available that spanned multiple adjacent catchments (data records were allocated to individual catchments).

Pinsky et al. 2009

Discovery applications

- General: SearchWorks
  - 7 million holdings at Stanford
  - Text search and facets
- GIS: GeoPortal
  - GeoBlacklight
  - Spatial search and visualization
  - Requires Spatial Data Infrastructure
- Each requires metadata workflows and repository



Collection
Hopkins Marine Station Collection

Format: Computer File
Summary:
This collection includes research data either produced by Hopkins Marine Station scientists or by other scientists who used Hopkins and the surrounding area as their study site. Data sets in this collection are in different formats, and each is presented with descriptive information that identifies the people and organizations responsible for its creation, related publications, and other relevant information. The collection was created to support the preservation of the research output of Hopkins Marine Station and to enable data sharing and re-use.

» Permanent URL for this item
» View 7 items in this collection online

Data Supplement for "Range-wide selection of catchments for Pacific salmon conservation."

Online
- purl.stanford.edu

Author/Creator: Pinsky, Malin L (Author)
Springmeyer, Dane B. (Author)
Goslin, Matthew N. (Author)
Augerot, Xanthippe (Author)
Language: English
Date created: 2009
Type of resource: Software, multimedia
Genre: Dataset
Format: Computer File
Abstract: Freshwater ecosystems are declining in quality globally, but a lack of data inhibits identification of areas valuable for conservation across national borders. We developed a biological measure of conservation value for six species of Pacific salmon (Oncorhynchus more
Preferred Citation: Pinsky, ML, Springmeyer, DB, Goslin, MN, Augerot, X (2009). Data Supplement for "Range-wide selection of catchments for Pacific salmon conservation." Stanford Digital Repository. Available at http://purl.stanford.edu/zc193vn8689.
Related Publication: Pinsky, M. L., Springmeyer, D. B., Goslin, M. N. and Augerot, X. (2009), Range-Wide Selection of Catchments for Pacific Salmon Conservation. Conservation Biology, 23: 680-691. doi: 10.1111/j.1523-1739.2008.01156.x. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1523-1739.2008.01156.x/full.
Collection: Hopkins Marine Station Collection
Related item: Wild Salmon Center
State of the Salmon
Abundance Estimates of the Pacific Salmon Conservation Assessment Database, 1978-2008
Subject: conservation planning
freshwater protected areas
landscape variables
northern Pacific
Oncorhynchus
salmon áreas dulceacuícolas protegidas
Oncorhynchus
Pacífico del norte
planificación de la conservación
salmón
variables de paisaje
Use and reproduction: User agrees that, where applicable, content will not be used to identify or to otherwise infringe the privacy or confidentiality rights of individuals. Content distributed via the Stanford Digital Repository may be subject to additional license and use restrictions applied by the depositor.

DRUID: zc193vn8689 | Librarian View

After self-deposit → Discovery in SearchWorks

Data Supplement for "Range-wide selection of catchments for Pacific salmon conservation."

Author/Creator:	Pinsky, Malin L (Author) Springmeyer, Dane B. (Author) Goslin, Matthew N. (Author) Augerot, Xanthippe (Author)
Language:	English
Date created:	2009
Type of resource:	Software, multimedia
Genre:	Dataset
Abstract:	Freshwater ecosystems are declining in quality globally, but a lack of data inhibits identification of areas valuable for conservation across national borders. We developed a biological measure of conservation value for six species of Pacific salmon (Oncorhynchus spp.) in catchments of the northern Pacific across Canada, China, Japan, Russia, and the United States. We based the measure on abundance and life-history richness and a model-based method that filled data gaps. Catchments with high conservation value ranged from Califor ... more »
Preferred Citation:	Pinsky, ML, Springmeyer, DB, Goslin, MN, Augerot, X (2009). Data Supplement for "Range-wide selection of catchments for Pacific salmon conservation." Stanford Digital Repository. Available at http://purl.stanford.edu/zc193vn8689.
Related Publication:	Pinsky, M. L., Springmeyer, D. B., Goslin, M. N. and Augerot, X. (2009), Range-Wide Selection of Catchments for Pacific Salmon Conservation. Conservation Biology, 23: 680-691. doi: 10.1111/j.1523-1739.2008.01156.x. Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1523-1739.2008.01156.x/full.
Contact:	malin.pinsky@gmail.com
Collection:	Hopkins Marine Station Collection
Related item:	Wild Salmon Center State of the Salmon Abundance Estimates of the Pacific Salmon Conservation Assessment Database, 1978-2008
Subject:	conservation planning freshwater protected areas landscape variables northern Pacific Oncorhynchus salmon áreas dulceacuícolas protegidas Oncorhynchus Pacífico del norte planificación de la conservación salmón variables de paisaje

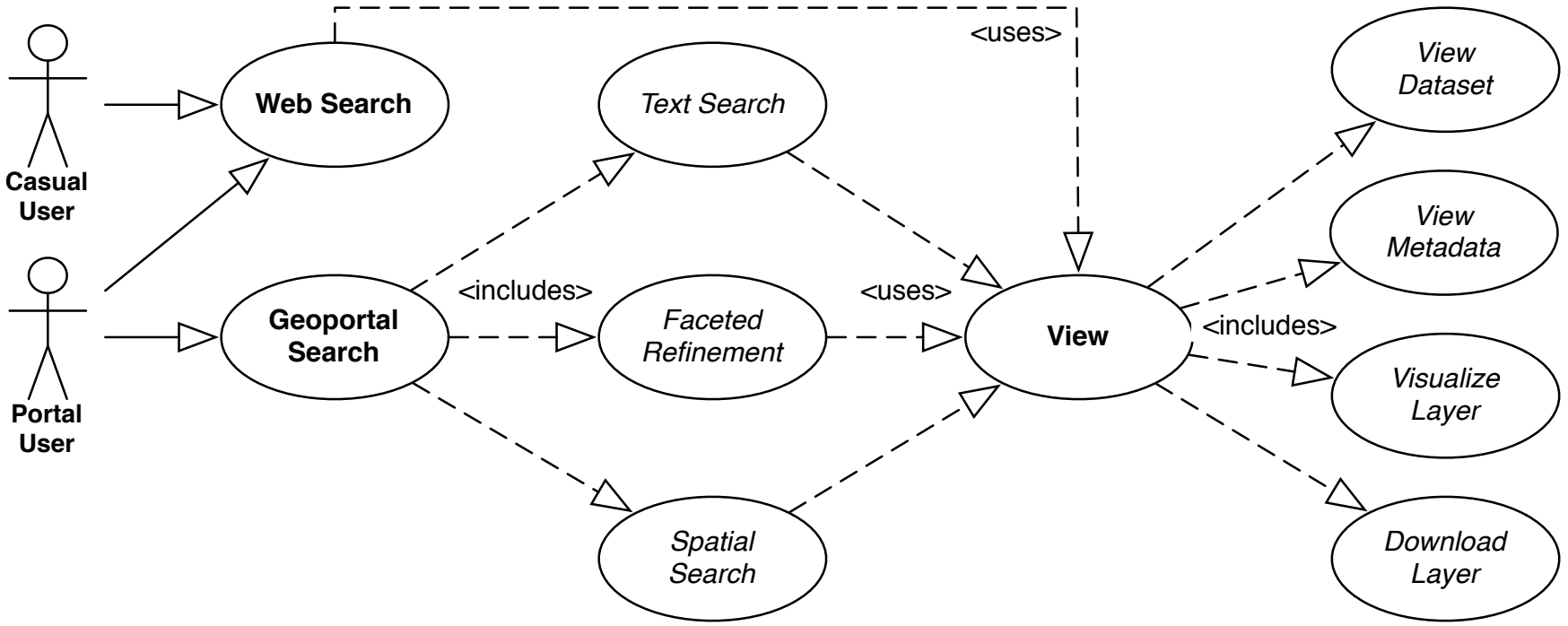
Citation-friendly **Persistent URL** for each digital object

PURL

http://purl.stanford.edu/zc193vn8689

File (Download)	Description
Pinsky_etal_AbundanceDatabase2008.csv (2 MB)	Abundance Database as published
Hydro1kshapefile.zip (6 MB)	Catchment ArcGIS shapefile to be used for georeferencing the Abundance and Run-timing databases (link CatchmentID in the database to NLEVEL5 in the shapefile). Shapefile includes river names.
Abundance_Database_Metadata.pdf (70 KB)	Describes the content of each field in the Abundance Database (Pinsky_etal_AbundanceDatabase2008.csv and Pinsky_etal_2009_AbundanceData-2010-09-29.csv)
RunTiming_Database_Metadata.pdf (70 KB)	Describes the content of each field in the Run Timing Database (Pinsky_etal_2009_RunTimingData-2010-09-29.csv)
Pinsky_etal_Sources.pdf (90 KB)	List of data sources for Abundance Database
Pinsky et al. 2009 Conservation Biology 23(3)-680-691 Supplemental Data_v2.pdf (40 KB)	Overview of this set of supplemental data files
Pinsky_etal_2009_AbundanceData-2010-09-29.csv (3 MB)	Revised Abundance Database; Removed duplicate sockeye abundance record from AK Peninsula (CatchmentID 15917)
Pinsky_etal_2009_RunTimingData-2010-09-29.csv (200 KB)	Run Timing Database

PURL provides delivery of self-deposited files



GeoBlacklight Discovery Use Cases

Discovery and delivery services in GeoBlacklight

[Demo]

Slippy map
Data inspection
Download
Share

Attribute	Value
fid	vv853br8653.394
objectid_1	373
nlevel5	15942.00000000000
studarea	1
areakm	1253.51517630000
cons_unit	SE BERING SEA INNER SHELF
name	Kvichak River, AK
lamazi_km	1246.6506289200
name2	Kvichak
lat	59.5147859104
long	-156.109041018
pa_km2	0E-17
pai_iv_km2	0E-17
country	United States

How to support GIS discovery?

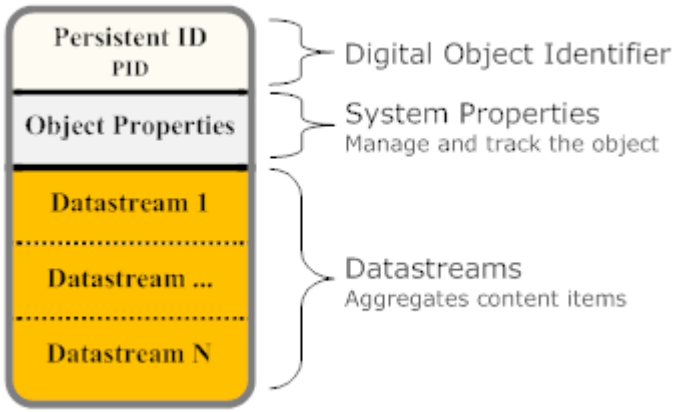
- Full-service repository for GIS data
  - Manage GIS data as *durable* digital library assets
  - Deliver vector, raster, and georectified maps via a spatial data infrastructure
  - Discover GIS assets via a variety of applications and contexts
  - Curate geospatial data via collaboratively maintained Hydra ecosystem

- At Stanford
  - *GeoHydra* + *Fedora* for metadata repository and workflow
  - *OpenGeo Suite* for delivery via Spatial Data Infrastructure
  - *GeoBlacklight* for discovery

Repository of Durable Digital Objects

Fedora

- PID
- Object props
- Data streams

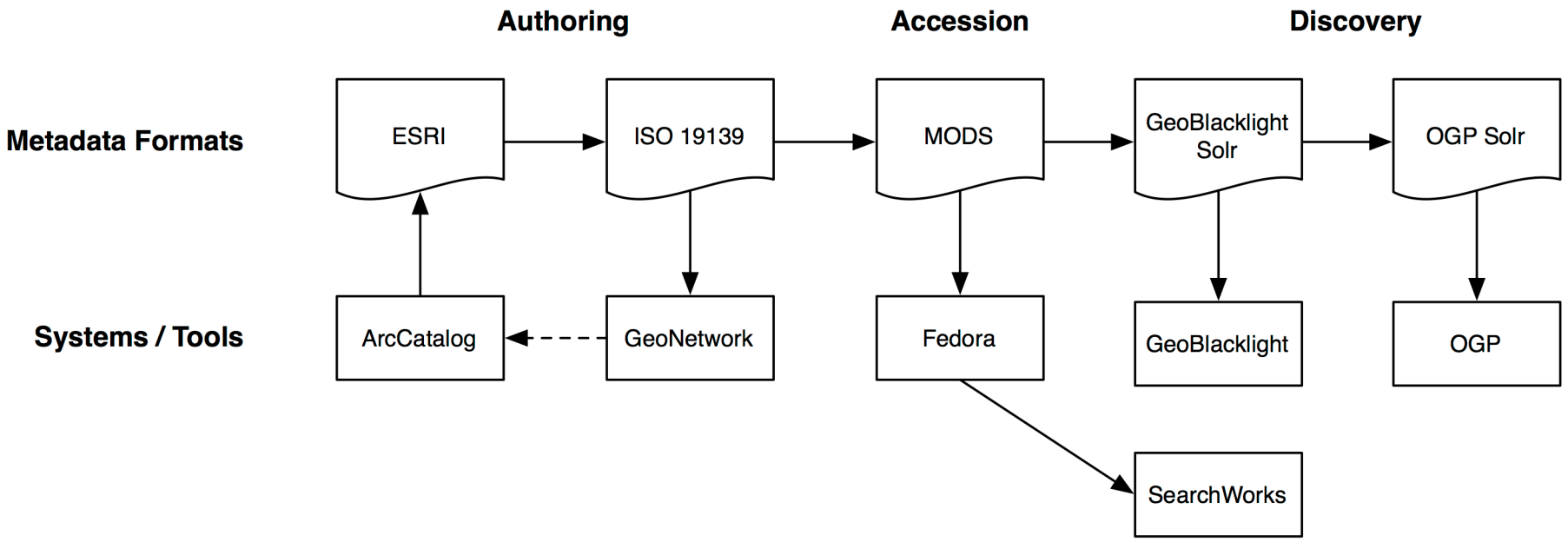


GeoHydra

Model, ingest, manage, and preserve geospatial data in Fedora

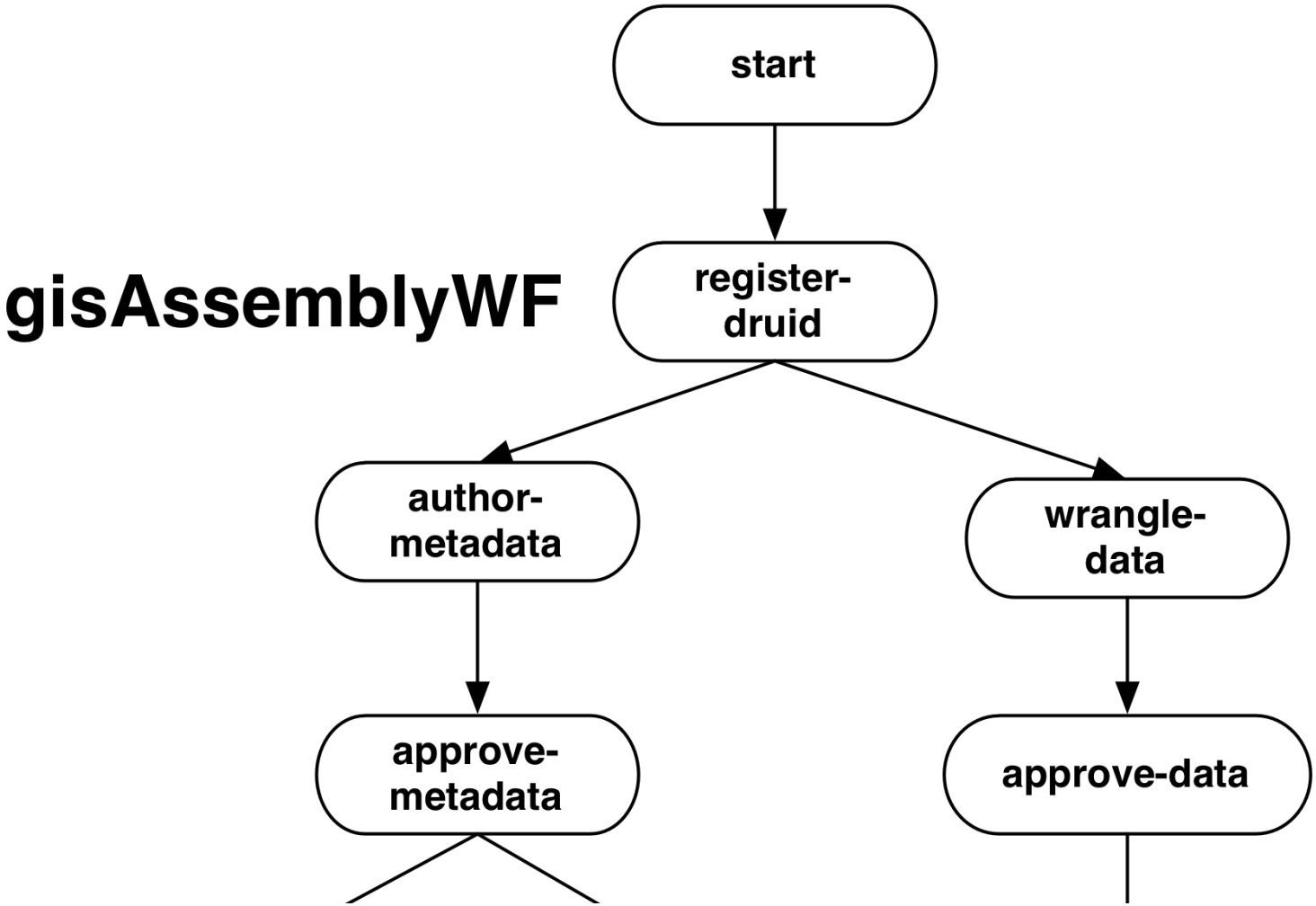
Digital Object Model for Repository

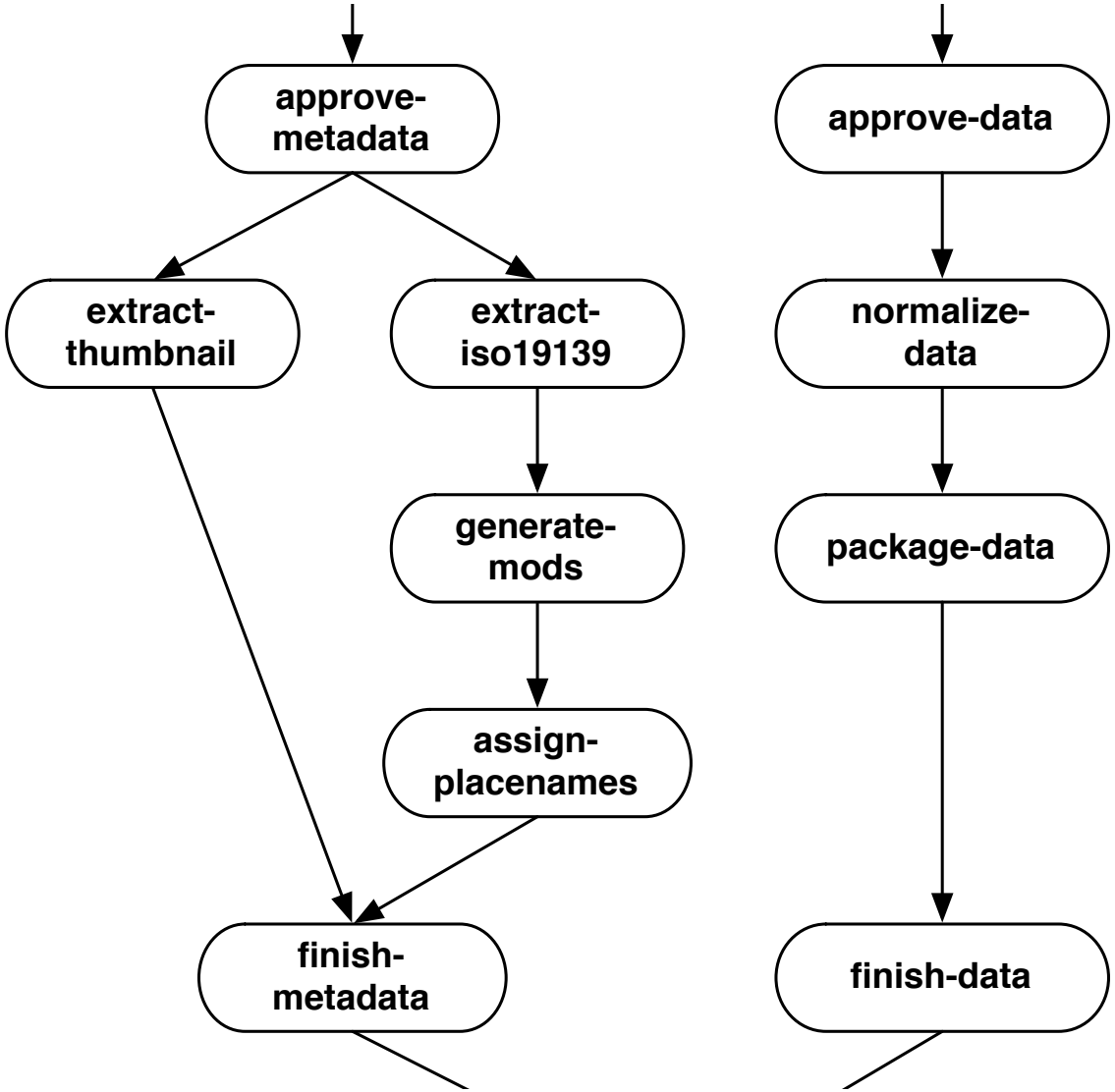
- Each GIS layer is a digital object
- Many GIS layers are a digital collection
- Datastreams Metadata
  - MODS “geo” extension
  - ISO 19139/19110 for GIS datasets
  - Linked data for place names (e.g., gazetteer)



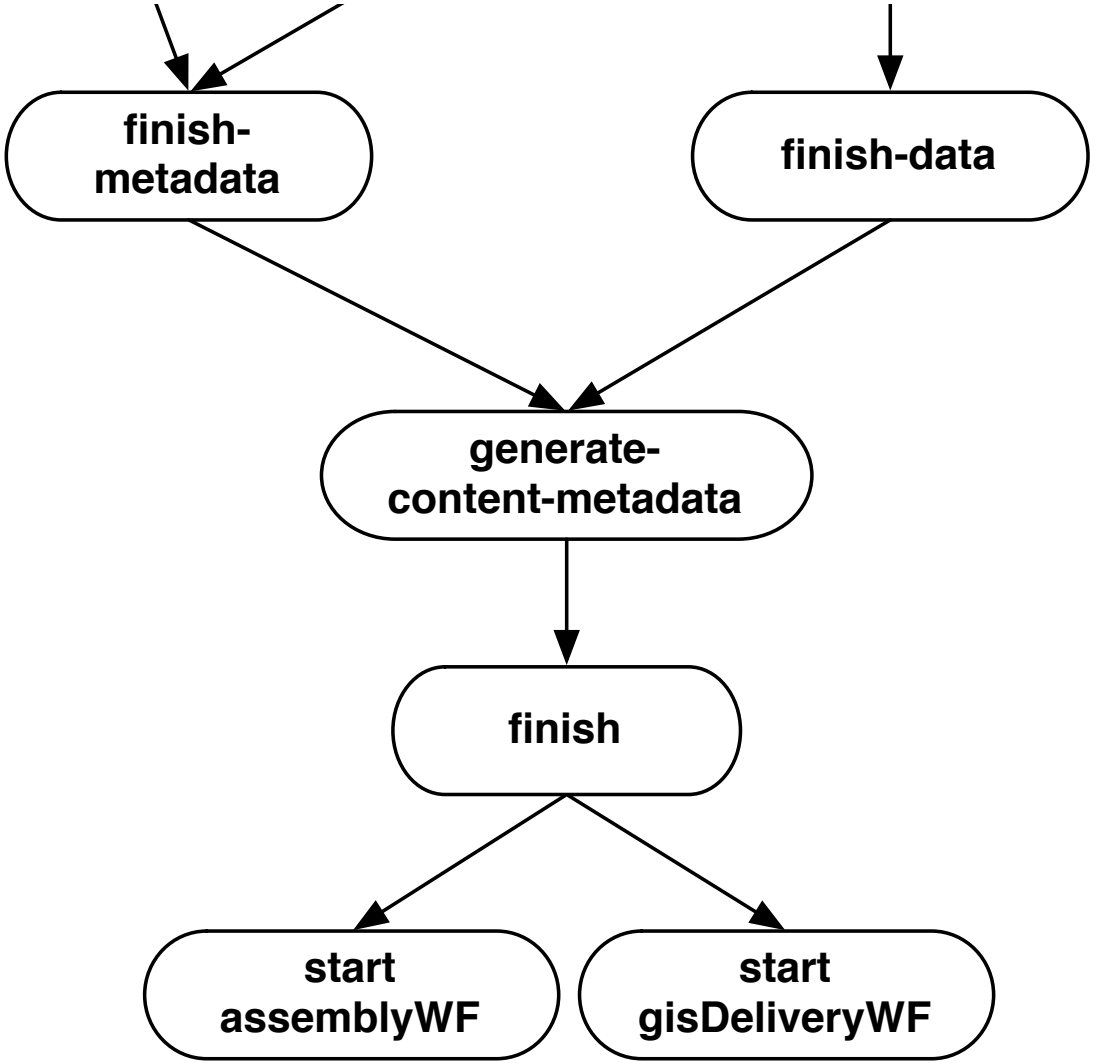
GIS metadata have many schemas

Workflows to Populate Repository





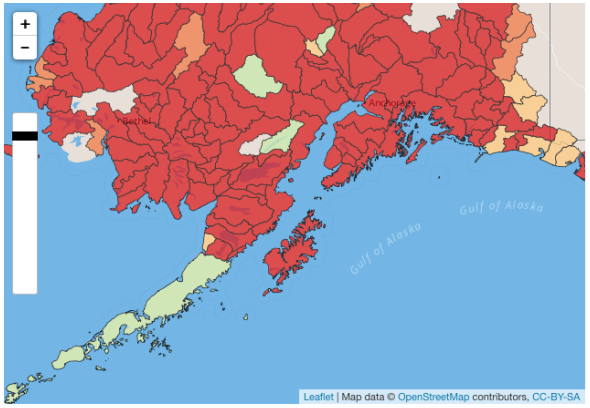
gisAssemblyWF



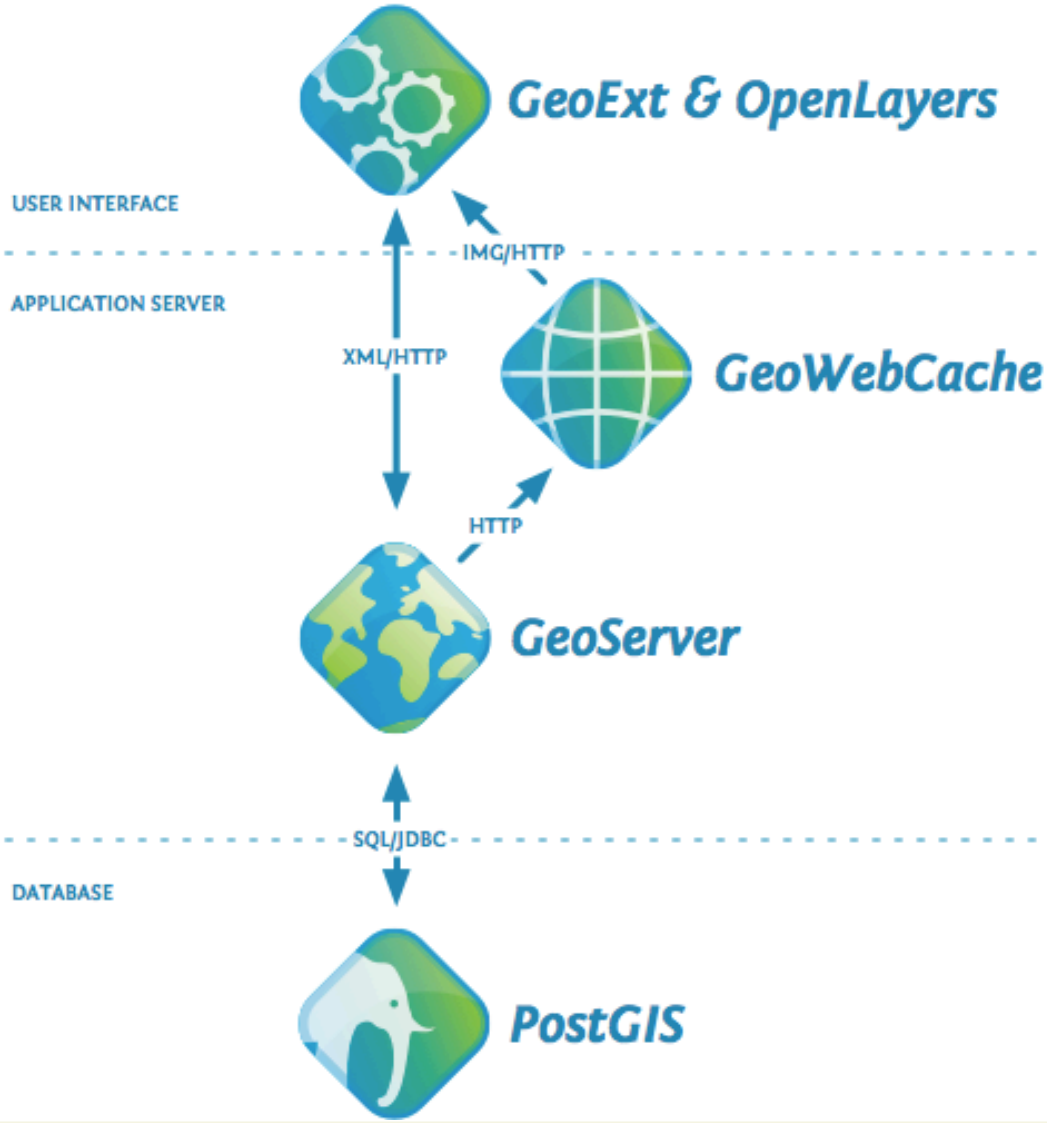
gisAssemblyWF

Wrangling Pinksy Data

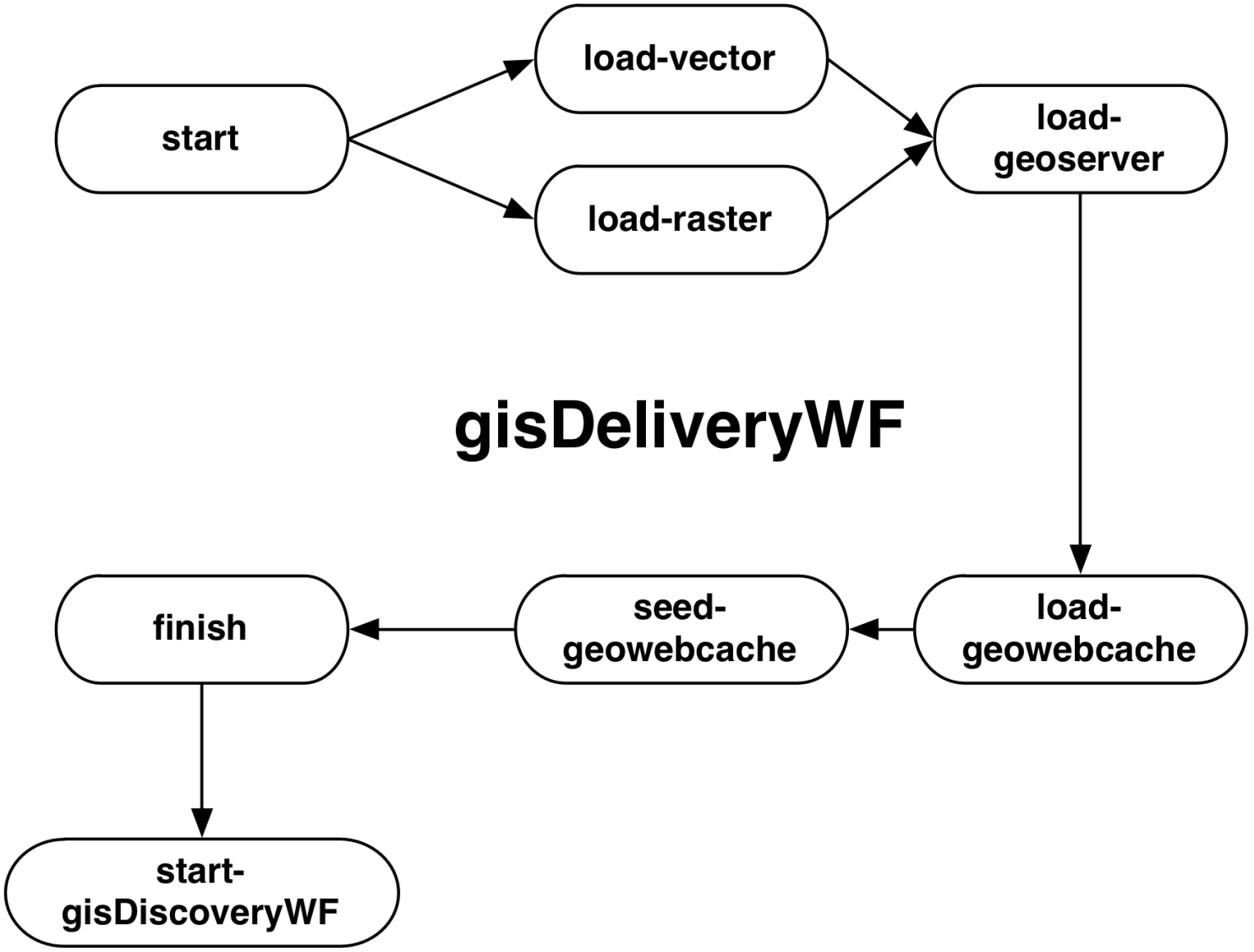
- GIS data (Shapefile) and study data (CSV) separated
  - Required table join, described in supplied metadata (PDF)
- Reproject data into standard WGS84
- Export joined GIS feature data into Shapefile format
- Author GIS metadata to bundle with Shapefile
  - Mostly transcribed from supplied metadata (PDF)
- Cartographic representation for 6 salmon species

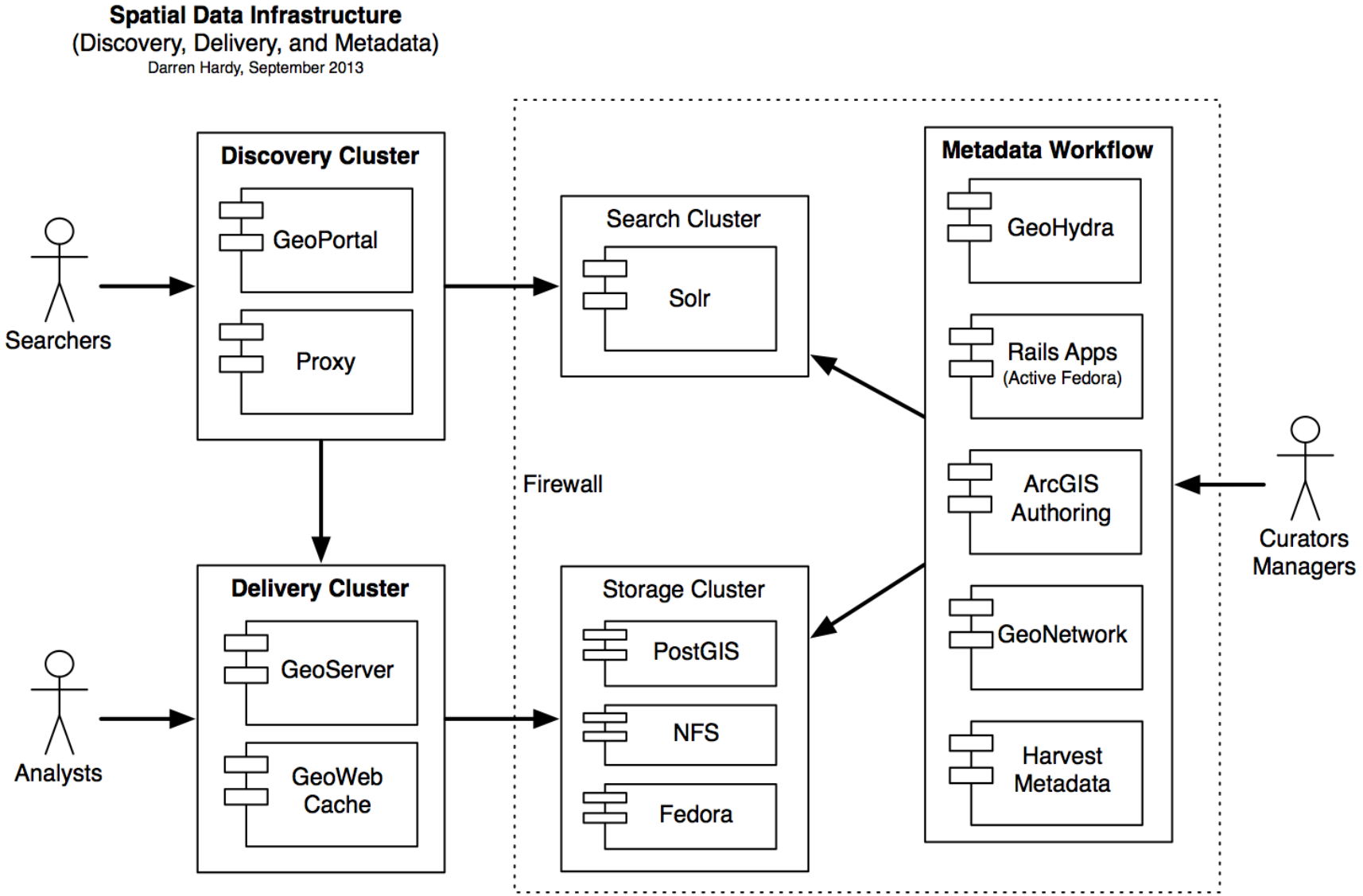


Attribute	Value
fid	vv853br8653.394
objectid_1	373
nlevel5	15942.00000000000
studarea	1
areakm	1253.51517630000
cons_unit	SE BERING SEA INNER SHELF
name	Kvichak River, AK
lamazi_km	1246.6506289200
name2	Kvichak
lat	59.5147859104
long	-156.109041018
pa_km2	0E-17
pai_iv_km2	0E-17
country	United States



Spatial Data Infrastructure using OpenGeo Suite by Boundless


start
load-vector
load-raster
load-geoserver
gisDeliveryWF
load-geowebcache
seed-geowebcache
finish
start-gisDiscoveryWF


Spatial Data Infrastructure
(Discovery, Delivery, and Metadata)
Darren Hardy, September 2013
Discovery Cluster
GeoPortal
Proxy
Searchers
Search Cluster
Solr
Metadata Workflow
GeoHydra
Rails Apps
(Active Fedora)
ArcGIS
Authoring
GeoNetwork
Harvest
Metadata
Firewall
Curators
Managers
Delivery Cluster
GeoServer
GeoWeb
Cache
Analysts
Storage Cluster
PostGIS
NFS
Fedora

Results

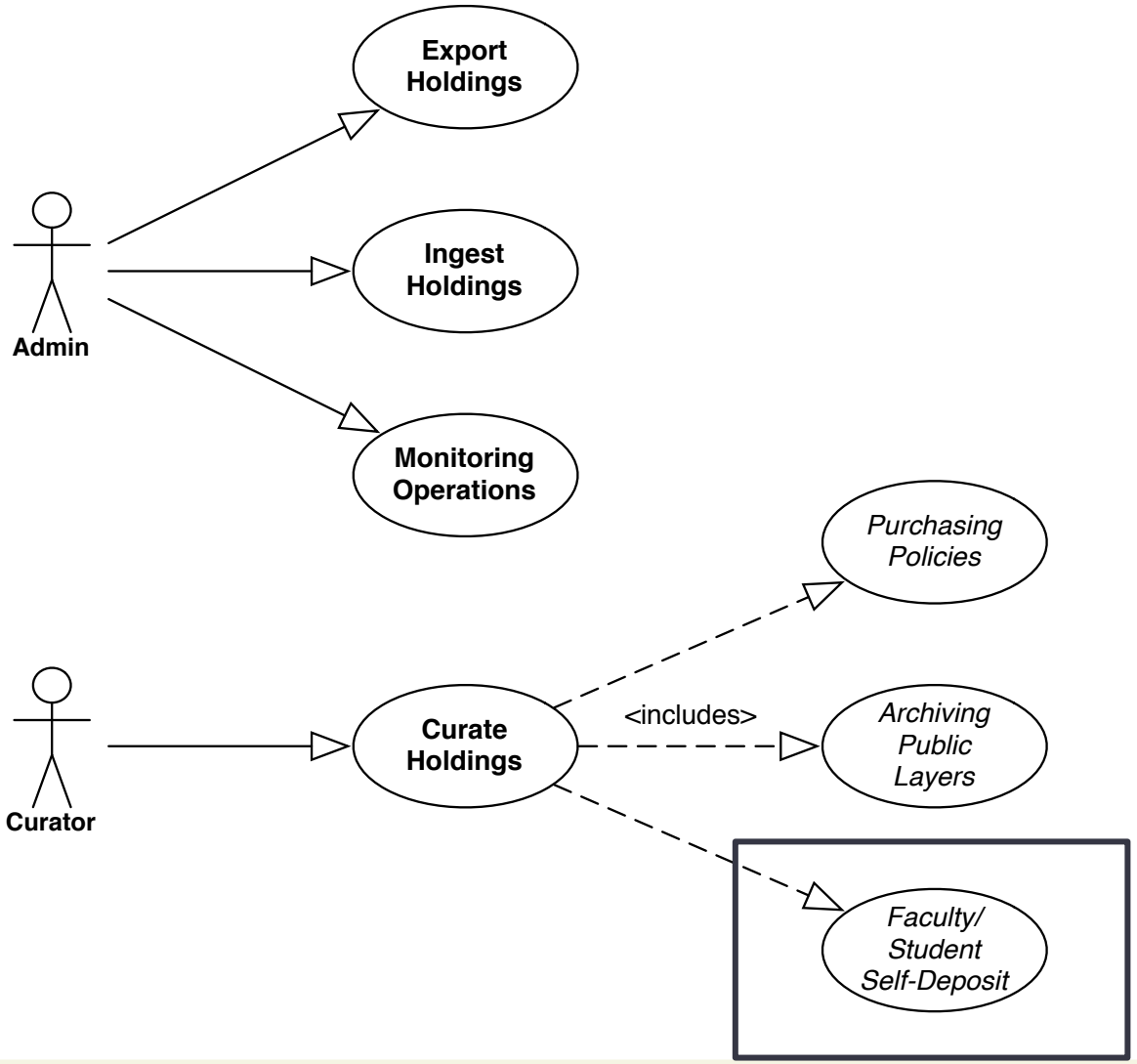
- Pinsky tries self-publish, site goes down
- Pinsky self-deposits GIS data and metadata (as-is)
- Stanford Digitial Repository
  - Provides PURL for data citation
- Workflows process GIS data and yield:
  - Discovery services via SearchWorks and GeoPortal
  - Delivery services via Spatial Data Infrastructure and simple file downloads
  - Long-term preservation

Summary

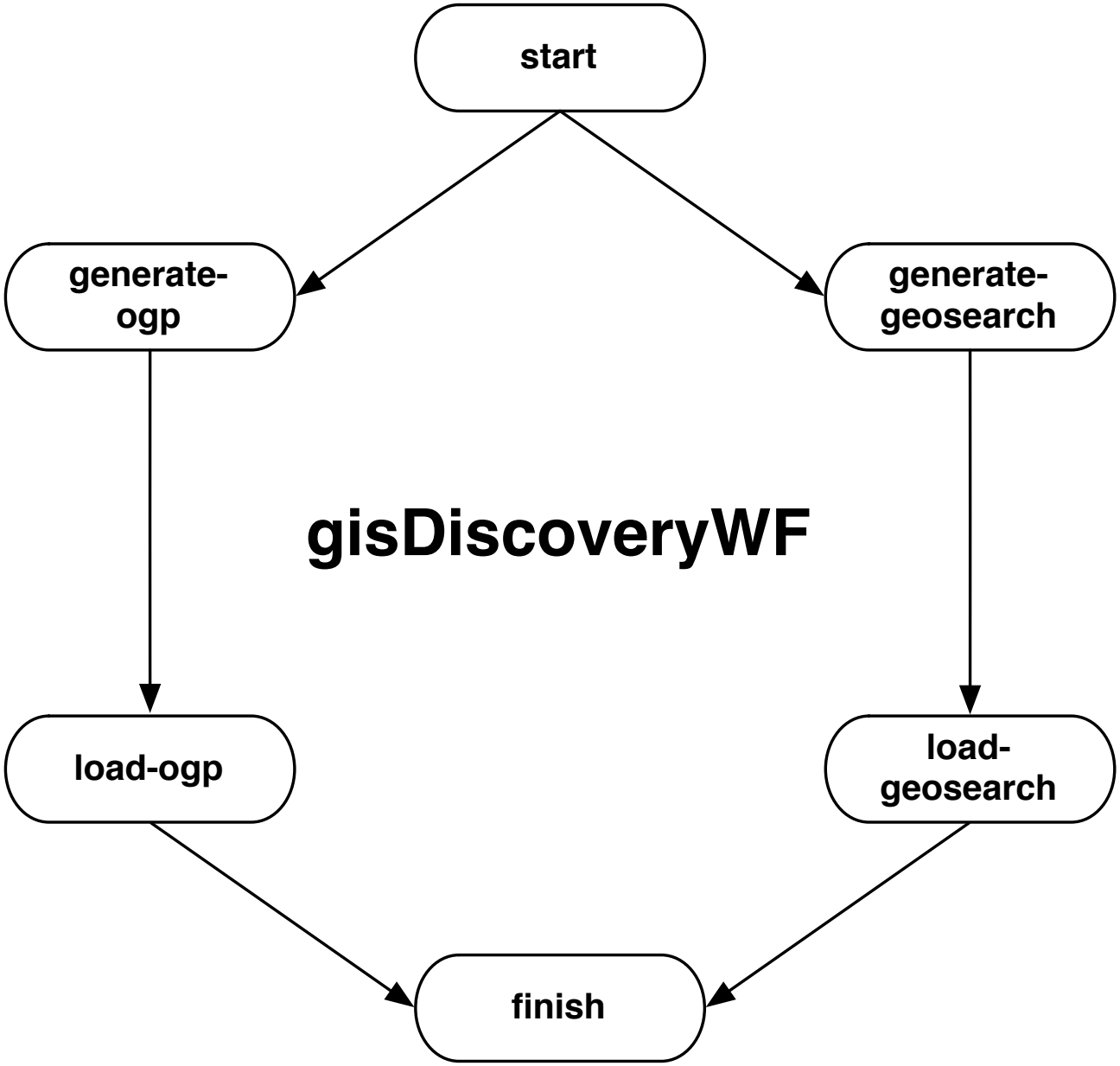
- Self-deposit into an institutional repository
  - Enables services for Discovery, Delivery, Preservation
  - Pinsky case study
- Manage GIS assets as *durable* digital library objects
  - But requires Spatial Data Infrastructure and GIS workflows

THANKS

Darren Hardy
Stanford University
Open Repositories 2014, Helsinki, Finland



Discovery Use Cases (Administration)


start
generate-
ogp
generate-
geosearch
gisDiscoveryWF
load-ogp
load-
geosearch
finish